Semaine d'Etude Mathématiques et Entreprises 5 : Sélection de variables statistiquement représentatives pour la production électrique photovoltaïque

Christophe Desmonts, Romain Bar, Marwa Hamza, Imen Chourabi, Papa-Abdulaye Faye

Semaine d'Etude Maths–Entreprises 5

11–15 février 2013, Ecole des Mines de Nancy

Sélection de variables statistiquement représentatives pour la production électrique photovoltaïque

Christophe Desmonts[a]
Romain Bar[a]
Marwa Hamza[a]
Imen Chourabi[b]
Papa-Abdulaye Faye[c]

[a] *Laboratoire de Mathématiques IECL, France*
[b] *Laboratoire de Mathématiques LMRS, France*
[c] *Laboratoire de Mathématiques Blaise Pascal, France*

Sujet proposé par :

Correspondants : Vincent Lefieux et Laurence Maillard-Teyssier

Résumé

La production électrique des panneaux photovoltaïques dépend de nombreux paramètres météorologiques : rayonnement du soleil, présence ou absence de nuages, température, ... La problématique que nous a soumise l'entreprise RTE et à laquelle nous réfléchissons dans ce document est de sélectionner les variables les plus influentes sur cette production au moyen d'une étude statistique, et de proposer un modèle descriptif de cette production qui adhère le mieux possible à la réalité. Dans cet objectif, nous faisons dans un premier temps un tour d'horizon des modèles statistiques existants. Nous étudions ensuite un modèle additif pour analyser les données fournies par RTE et effectuer une première sélection de variables grâce au modèle GAM. Enfin, on reprend cette étude avec le modèle MARS dans l'objectif de pouvoir regrouper des variables entre elles pour pouvoir transformer notre modèle additif très restrictif en un modèle plus adapté à la situation considérée.

Mots clés : Modèle GAM, modèle MARS.

Numéro de publication : SEME005-2013-02-D

1 Introduction

La transition énergétique fait l'objet d'un vaste débat national. Quels que soient les scénarios retenus, RTE (Réseau de Transport d'Electricité) accompagnera cette transition et en particulier la croissance des énergies renouvelables.

L'entreprise RTE conduit ses activités de recherche dans l'objectif d'améliorer la sécurité d'alimentation tout en minimisant l'impact environnemental. Elle veille à une gestion optimale du flux d'électricité, et doit à ce titre connaître la production photovoltaïque. Ceci passe par l'étude des facteurs qui agissent sur cette production : température, rayonnement, vent, nébulosité, ... Ces facteurs étant pléthoriques, il est nécessaire de pouvoir déterminer quels sont ceux qui conditionnent le plus la production photovoltaïque.

Du point de vue mathématique, notons P, T, V, R et N respectivement la production, la température, le vent, le rayonnement et la nébulosité. La problématique est de déterminer une fonction f telle que

$$P = f(T, V, R, N). \tag{1}$$

Notre objectif est de préciser quelles sont les variables prépondérantes dans cette relation. Un premier point de vue est de chercher la fonction f sous une forme additive, c'est à dire en réécrivant (1) comme suit

$$P = f_1(T) + f_2(V) + f_3(R) + f_4(N). \tag{2}$$

Ce modèle est a priori déjà très restrictif, et pose une question naturelle ; y a t'il des interactions entre les différentes variables ? C'est à dire peuvons-nous exprimer la production comme suit

$$P = f_1(T) + f_2(V) + ... + g_1(T, V) + g_2(T, R) + g_3(T, V, R) + ... \ ? \tag{3}$$

Le but de ce rapport est de présenter les différentes pistes que nous avons étudiées pour répondre à cette question. Dans la première partie nous expliquons brièvement le principe de la modélisation non paramétrique. La deuxième partie présente la méthode de modélisation statistique GAM avec une interprétation des résultats obtenus en tenant compte de nos données physiques.
Enfin dans la troisième section nous rappelons la méthode de modélisation MARS, nous donnons une prédiction de la production et nous précisons, suivant les algorithmes MARS, les variables prépondérantes. Finallement, nous donnons la conclusion et les perspectives.

2 Méthodes existantes

Le but est de modéliser la dépendance d'une variable réponse Y à partir de variables explicatives $X^1, \ldots, X^p$.

On suppose que l'on modélise cette relation de la manière suivante :

$$Y = f(X^1, \ldots, X^p) + \epsilon,$$

où ϵ, représentant l'aléa, est supposé centré.

Le but est de déterminer la fonction lien f grâce aux données (observations) à notre disposition pour pouvoir prédire les futures sorties y à partir de la connaisances des $x^i, i = 1, \ldots, p$.

Pour les simulations, on choisit un échantillon d'apprentissage variant de 1000 à 10000 données et on réalise une prédiction sur 1000 autres données. L'erreur globale est calculée de la manière suivante : $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{1000}\sum_{i=1}^{1000}(y_i - \hat{y}_i)^2$. Les résultats sont donnés dans la dernière section de la partie. La sélection de variables statistiquement représentatives sera également discutée dans cette dernière section.

2.1 L'approche non-paramétrique :

On impose moins la structure des prédicteurs, on suppose qu'ils appartiennent à une certaine classe de fonctions, par exemple aux splines d'ordre q : polynômes par morceaux de degré inférieur ou égal à q.

$$\hat{f}(x) = \hat{g}(x, (\hat{a_j})), \qquad \hat{g}(x) = \underset{g}{arg\,min} \sum_{i=1}^{n}[y_i - g(x, (a_j))]^2.$$

Malheureusement, en grande dimension, le nombre de paramètres a_j à estimer devient trop important en pratique.

Une idée intéressante est alors de tenter de réduire la dimensionalité pour pallier ce problème.

On propose alors :

$$\hat{f}(x) = \sum_{j=1}^{J} \hat{g}_j(z^j, a)$$

où z^j est un sous-ensemble de $(x^1, \ldots, x^p)$.

$$\hat{g_j}(z_j) = \underset{g_j}{arg\,min} \sum_{i=1}^{n}[y_i - \sum_{j=1}^{J} g_j(z_j)]^2.$$

Ainsi, la fonction f $n-$dimensionnelle est approchée par J fonctions dont l'argument est de plus faible dimension.

Remarque : une variable peut intervenir dans plusieurs des sous-ensembles $z_j, j = 1, \ldots, p$.

Le modèle le plus connu est celui où on pousse la démarche au maximum en travaillant avec des fonctions unidimensionnelles i.e. :

$$\hat{f}(x) = \sum_{j=1}^{p} \hat{g}_j(x^j, a).$$

C'est le modèle GAM (Generalized Additive Model) qui sera davantage étudié dans la partie suivante.

Les techniques de partitionnement récursif :

Idée : on partitionne l'espace $D \subset \mathbb{R}^p$ en régions disjointes et, pour chaque région, on calcule un estimateur ce qui confère à ces techniques un aspect local avantageux :

si $x \in R_m$, alors

$$\hat{f}(x) = \hat{g_m}(x).$$

Les régions R_m sont disjointes et forment une partition du domaine D.

Les arbres de régression font partie de ces techniques : si $x \in R_m$, alors

$$\hat{f}(x) = \hat{g_m}(x) = \frac{1}{card(R_m)} \sum_{x_i \in R_m} y_i$$

Le prédicteur global devient alors : $\sum_{m=1}^{M} \mathbf{1}_{\{x \in R_m\}} \hat{g_m}(x)$, constant par morceaux.

Principe de construction :

On part de l'espace D tout entier.

1) Etape "forward" :

A chaque étape, on effectue un partitionnement optimal (en 2) de chaque région.

Par exemple, pour les arbres, on minimise la variance des régions "filles" :

$$(j^*, d^*) = \underset{j,d}{argmin}\Big(\sum_{i:x_i \in R_1} (y_i - y_{\bar{R}_1})^2 + \sum_{i:x_i \in R_2} (y_i - y_{\bar{R}_2})^2 \Big).$$

2) Etape "backward" :

On recombine les sous-régions jusqu'à obtenir une représentation optimale au sens d'un critère général réalisant un compromis entre la qualité d'ajustement et le nombre de régions.

Pour les arbres, cette étape s'appelle l'élagage, on minimise le critère :

$$C_\lambda(M) = \sum_{m=1}^{M} (y_i - y_{\bar{R}_m})^2 + \lambda \cdot M$$

Le problème est que le prédicteur n'est pas continu ce qui peut s'avérer ennuyeux si la fonction sous-jacente f l'est.

Une idée est alors de lisser la fonction indicatrice pour gagner en régularité.

Dans la méthode MARS (Multivariate Adaptive Regression Splines) par exemple, la fonction indicatrice des méthodes par arbre est approchée par des fonctions splines.

Les méthodes d'ensemble.

Idée : Construire une collection de prédicteurs (en introduisant de l'aléa dans l'échantillon) et agréger l'ensemble de leurs prédictions. (en faisant la moyenne par exemple)

exemple : les forêts aléatoires.

Chaque prédicteur est calculé avec une méthode d'arbres sur un échantillon bootstrap de l'échantillon initial.

A chaque étape de partitionnement de l'arbre, on ne garde qu'un sous ensemble de variables, tirées au hasard (de manière uniforme par exemple).

Remarque : on pourrait très bien imaginer une version ensembliste de la méthode MARS basée sur le même principe.

3 Résultats

Variance résiduelle :

nbre d'individus dans l'échantillon d'apprentissage / méthode	1000	2000	3000	4000	5000	6000	8000	10000
régression linéaire sur toutes les var. explicatives	23.07	6.94	3.91	2.72	2.54	2.11	1.60	1.44
AFM + reg. lin. sur les 3 premiers facteurs	20.39	6.26	3.89	2.88	3.37	2.77	2.05	1.62
régression PLS	447.43	130.65	75.56	53.21	65.06	53.97	40.88	43.58
arbre de régression	26.76	8.92	5.15	3.83	2.93	2.69	1.86	1.51

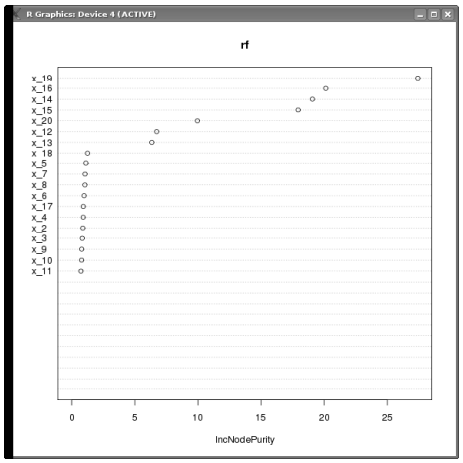
Sélection de variables :

i) arbre de régression :

```
node), split, n, deviance, yval
      * denotes terminal node

 1) root 499 19.72027000 0.103252600
   2) x_12< 210.8 404  2.07876300 0.022351340
     4) x_19< 45 373  0.27862660 0.007665603 *
     5) x_19>=45 31  0.75175410 0.199053900
      10) x_20< 94.5 24  0.19670280 0.143648300 *
      11) x_20>=94.5 7  0.22877830 0.389015700 *
   3) x_12>=210.8 95  3.75257100 0.447295900
     6) x_19< 119 55  1.64225400 0.347140900
      12) x_14< 184.1 16  0.05975628 0.189138800 *
      13) x_14>=184.1 39  1.01919200 0.411962300 *
     7) x_19>=119 40  0.80001530 0.585009000 *
```

ii) méthode des forêts aléatoires :



iii) méthode mars :

```
Selected 15 of 17 terms, and 9 of 19 predictors
Importance: x_19, x_14, x_4, x_20, x_7, x_17, x_18, x_15, x_16, x_2-unused, x_3-unused, x_5-unused, x_6-unused, x_8-unused, x_9-unused, x_10-unused, x_11-unused, ...
Number of terms at each degree of interaction: 1 14 (additive model)
GCV 0.008759249    RSS 17.00539    GRSq 0.8542297    RSq 0.8582867
```

Quelle que soit la méthode, on constate que les variables 19 : “Rayonnement à la station 1”, 14 : “Rayonnement au point grille 3” et 20 : “Rayonnement la station 2” semblent être les plus statistiquement représentatives.
Il est remarquable que ces 3 variables soient liées au phénomène de rayonnement ; ceci nous laisse penser qu’il s’agit du phénomène prépondérant pour prédire la production photovoltaïque.

4 Le modèle GAM

L'objectif de cette section est d'utiliser une méthode statistique déjà existante pour calculer les fonctions f_i du modèle additif, et d'essayer de ne retenir que les paramètres qui influent le plus. L'idée est ensuite d'essayer de regrouper certains paramètres pour se défaire de la restrictivité du modèle purement additif et pour se rapprocher de la réalité.

4.1 Calcul des paramètres prépondérants et prédictions

On utilise GAM (Generalized Additive Model) pour déterminer quelles sont les données les plus importantes. Tous les codes de cette section sont utilisés par le logiciel R. Le fichier `Tableaudedonnees.csv` contient le tableau des 5000 premiers relevés (du 3 janvier 2009 au 30 juillet 2009) de la production en fonction des relevés des 19 paramètres :

```
x<-read.table("Tableaudedonnees.csv",header=T,dec=",",sep=";")
library(gam)
REG=gam(charge~s(temp1)+s(temp2)+s(temp3)+s(temp4)+s(temp5)
+s(vent1)+s(vent2)+s(vent3)+s(vent4)+s(vent5)+s(ray1)+s(ray2)
+s(ray3)+s(ray4)+s(ray5)+s(tempstat0)+s(nebu0)+s(raystat1)
+s(raystat2),data=x)
summary(REG)
```

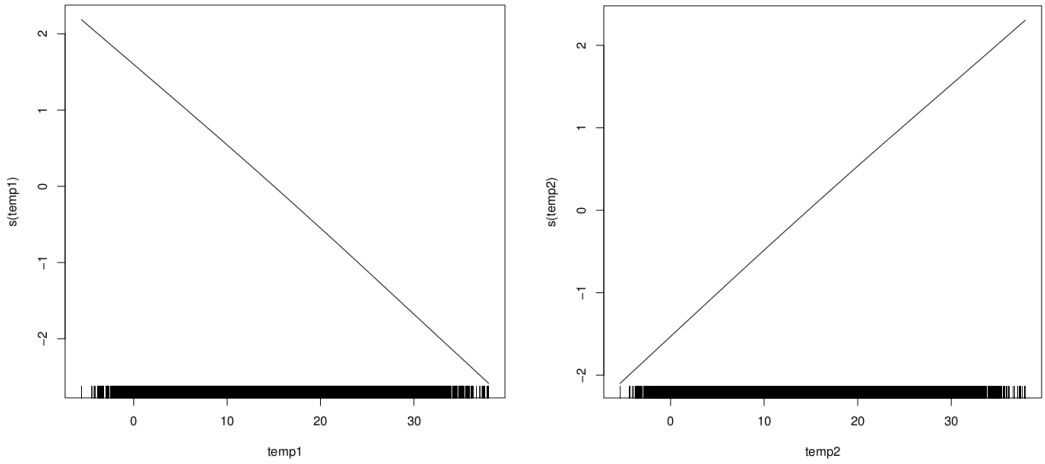
La commande `summary` permet de renvoyer toute une série de grandeurs sur le tableau de données comme la variance, la moyenne etc ... La réponse de R contient également des informations sur l'importance des variables, manifestée par la donnée de p-valeurs (dernière colonne) :

```
             Df Npar Df  Npar F     Pr(F)
(Intercept)   1
s(temp1)      1       3 109.163 < 2.2e-16 ***
s(temp2)      1       3  72.926 < 2.2e-16 ***
s(temp3)      1       3   6.090 0.0003887 ***
                .........
s(vent2)      1       3   3.490 0.0149933 *
s(vent3)      1       3   0.978 0.4019091
s(vent4)      1       3   0.913 0.4334614
                .........
s(ray3)       1       3   7.384 6.101e-05 ***
s(ray4)       1       3   7.600 4.474e-05 ***
s(ray5)       1       3   7.694 3.906e-05 ***
s(tempstat0)  1       3   2.721 0.0427745 *
s(nebu0)      1       3   8.328 1.569e-05 ***
s(raystat1)   1       3  23.489 3.664e-15 ***
s(raystat2)   1       3  21.027 1.367e-13 ***
```

Plus la p-valeur est petite, plus la variable a une grande probabilité d'être importante ; d'après ces relevés, les variables les plus importantes sont donc temp1, temp2, temp3,

temp4, temp5, ray3, ray4, ray5, nebu0, raystat1 et raystat2. Notons que la présence ou l'absence de vent influe très peu sur la production photovoltaïque.

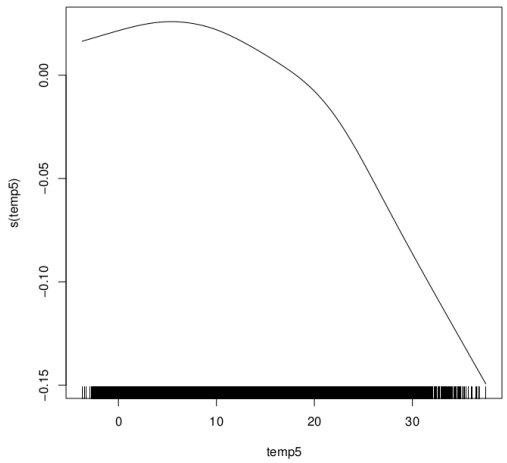
On relance le même code, puis on trace les courbes des fonctions grâce à la commande `plot(REG)`. On constate alors un problème, visible sur les courbes de temp1, temp2, temp3, temp4,ray4 et ray5 :

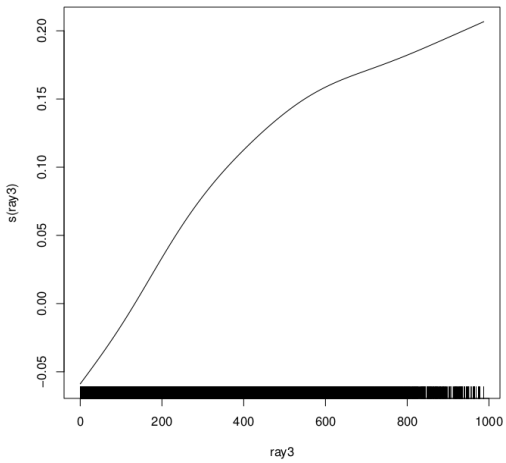


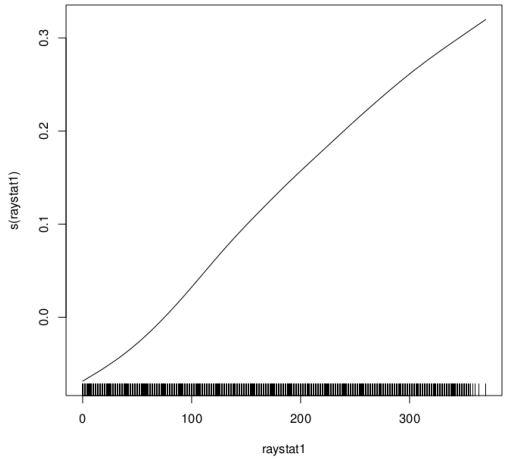
Les courbes de temp1 et temp2 (respectivement de temp3 et temp4 et de ray4 et ray5) sont croisées. Noter que ce sont à chaque fois sur des relevés donnés par des stations proches l'une de l'autre. Analytiquement, ceci se traduit ainsi :

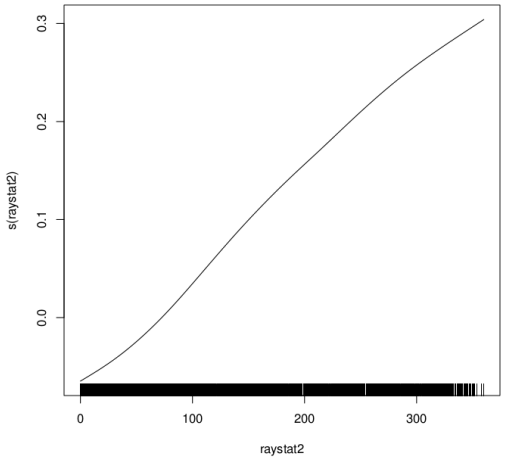
$$\begin{aligned} P \;=\; & \underbrace{f_1(\text{temp1}) + f_2(\text{temp2})}_{\simeq 0} + \underbrace{f_3(\text{temp3}) + f_4(\text{temp4})}_{\simeq 0} + f_5(\text{temp5}) \\ & + f_6(\text{ray3}) + \underbrace{f_7(\text{ray4}) + f_8(\text{ray5})}_{\simeq 0} + f_9(\text{nebu0}) + f_{10}(\text{raystat1}) \\ & + f_{11}(\text{raystat2}). \end{aligned}$$

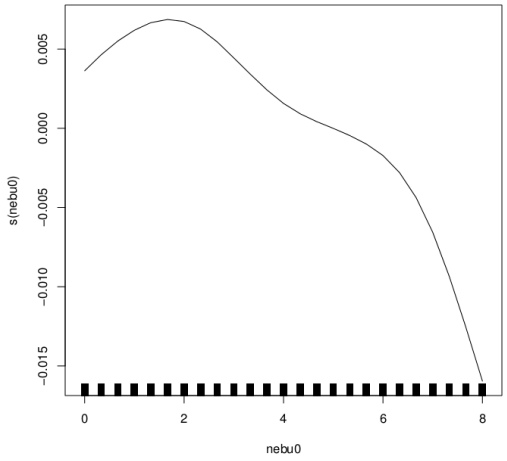
Le logiciel R aurait ainsi pu tout aussi bien prendre des fonctions quasiment identiquement nulle pour les fonctions représentant l'impact de temp1, temp3, temp4, ray4 et ray5. Ce ne sont donc pas *a priori* des variables importantes ; en les enlevant une à une de la liste des paramètres pris en compte, on constate que ceci se vérifie. En définitive, il ne reste donc que les variables temp5, ray3, nebu0, raystat1 et raystat2. Remarquons que les courbes restent très similaires si on remplace temp$i$ par temp$j$ ou ray$i$ par ray$j$, où i et j désignent des stations proches l'une de l'autre.


s(temp5)
0.00
−0.05
−0.10
−0.15
0
10
20
30
temp5


s(ray3)
0.20
0.15
0.10
0.05
0.00
−0.05
0
200
400
600
800
1000
ray3


s(raystat1)
0.3
0.2
0.1
0.0
0
100
200
300
raystat1


s(raystat2)
0.3
0.2
0.1
0.0
0
100
200
300
raystat2


s(nebu0)
0.005
0.000
−0.005
−0.010
−0.015
0
2
4
6
8
nebu0

Interprétation des courbes :

- Température au point grille 5 : on constate que la production photovoltaïque est optimale pour une température d'environ 8°C. On voit également que des températures trop élevées nuisent à la production dans des proportions assez importante (-10% à 30°C ; rappelons que la production mesurée est en fait un facteur de charge, compris entre 0 et 1).
- Rayonnement au point grille 3 et aux stations : conformément à ce que l'on pouvait attendre, le rayonnement a un très gros impact sur la production photovoltaïque. Plus précisément, ce sont les rayonnements mesurés aux stations qui jouent le plus, et de façon similaire.
- Nébulosité : on voit sur l'échelle de l'axe des ordonnées que la nébulosité (c'est à dire la présence de nuages) influe très peu sur la production. Il semble tout de même qu'elle y contribue de manière légèrement positive lorsqu'il y a un peu de nuages. Ce résultat peut sembler étonnant : la logique voudrait que la présence de nuage influe beaucoup sur la production, et que celle-ci soit optimale lorsqu'il n'y a aucun nuage. Ce qui peut sembler être un paradoxe peut s'expliquer par le fait que nous n'avons pas pris en compte le phénomène des saisons dans le choix de nos données. En toute rigueur, il faudrait sélectionner des données saisonnières et tracer les courbes pour chaque saison.

4.2 Evaluation du modèle

Pour évaluer la finesse du modèle

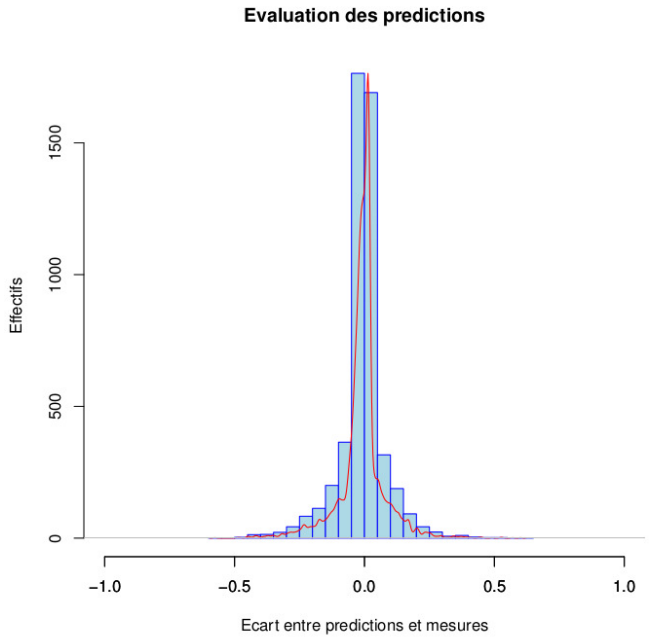
$$P = f_1(\text{temp5}) + f_2(\text{ray3}) + f_3(\text{nebu0}) + f_4(\text{raystat1}) + f_5(\text{raystat2})$$

obtenu, on va l'utiliser pour faire des prédictions sur les 5000 dernières données du tableau (du 6 juin 2010 au 31 décembre 2010) :

```
z<-read.table("5000dernieresvaleurs.csv",header=T,
dec=",",sep=";")
Y=predict(REG,z)
```

Le vecteur Y contient les 5000 valeurs théoriques de la production, calculées grâce au modèle à partir des 5000 dernières valeurs de temp5, ray3, nebu0, raystat1 et raystat2 du tableau. Reste à les comparer aux valeurs effectivement relevées. On trace pour cela un histogramme de dispersion de l'écart quadratique entre prédictions et relevés, ainsi que la densité de cet écart.

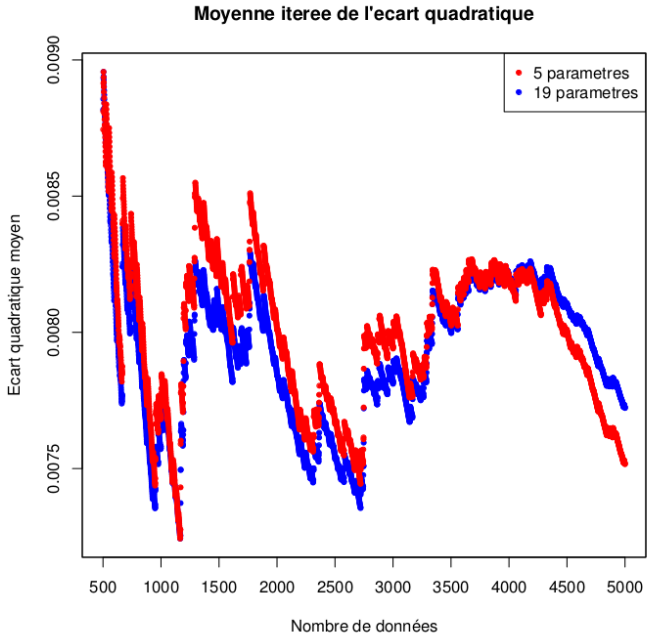
```
D=Y-z[,1]
hist(D,xlim = c(-1,1),main="Evaluation des predictions",
xlab="Ecart entre predictions et mesures",
ylab="Effectifs",col="lightblue",border="blue",breaks=20)
den<-density(D)
par(new = TRUE)
plot(den,xlim = c(-1,1),main="",xlab="",ylab="",col = "red",
yaxt="n",bty="n")
graphics.off()
```



On compare maintenant la qualité des prédictions obtenues avec ces 5 paramètres par rapport à celles qu'on obtient en conservant les 19 paramètres. Pour ce faire, on compare sur un même graphe les courbes de la moyenne de l'écart entre prédiction et relevé en fonction du nombre de données considérées.

```
x<-read.table("Tableaudedonnees.csv",header=T,dec=",",sep=";")
library(gam)
REG1=gam(charge~s(temp1)+s(temp2)+s(temp3)+s(temp4)+s(temp5)
+s(vent1)+s(vent2)+s(vent3)+s(vent4)+s(vent5)+s(ray1)+s(ray2)
+s(ray3)+s(ray4)+s(ray5)+s(tempstat0)+s(nebu0)+s(raystat1)
+s(raystat2),data=x)
REG2=gam(charge~s(temp5)+s(ray3)+s(nebu0)+s(raystat1)+s(raystat2),
data=x)
z<-read.table("5000dernieresvaleurs.csv",header=T,dec=",",sep=";")
Y1=predict(REG1,z)
Y2=predict(REG2,z)
D1=Y1-z[,1]
D2=Y2-z[,1]
D1<-D1*D1
D2<-D2*D2
w<-1:5000
w<-1/w
p1=cumsum(D1)
p2=cumsum(D2)
K1=p1*w
K2=p2*w
K1<-K1[-c(1:500)]
```

```
K2<-K2[-c(1:500)]
pdf("Comparaison.pdf")
plot(K1,pch=20,col="blue",main="Moyenne iteree de l'ecart
quadratique",xlab="Nombre de données",ylab="Ecart quadratique
moyen",xaxt="n")
par(new=T)
plot(K2,pch=20,col="red",main="",xlab="",ylab="",yaxt="n",xaxt="n")
axis(1,at=seq(0,4500,by=500),labels=seq(500,5000,by=500))
legend("topright", legend = c("5 parametres", "19 parametres"),
col = c("red", "blue"),pch = 20)
dev.off()
```



On constate que les prédictions faites en ne considérant que les 5 variables considérées comme les plus importantes sont plus précises que celles faites à partir des 19 variables, ce qui est tout à fait satisfaisant.

On peut utiliser cette technique de prédictions pour tester l'importance des variables :
- On choisit la variable dont on veut tester l'importance (par exemple temp1) ;
- On mélange aléatoirement toutes les données de la colonne de valeurs de temp1 ;
- On lance GAM avec le nouveau jeu de données ;
- On compare les prédictions obtenues à partir des données mélangées avec celles obtenues avec les données de départ : plus l'écart obtenu est grand, plus la variable est importante.

4.3 Amélioration par le modèle physique

Le choix d'un modèle additif est très restrictif a priori ; il n'y a techniquement aucune raison que la production s'écrive comme une somme de fonctions d'une seule variable. Pour

se rapprocher de la réalité, il faut regrouper intelligemment certaines de ces variables pour créer de nouvelles listes de données, et obtenir une fonction plus réaliste. Par exemple, on regroupe les données temp1 et ray1 en une nouvelle colonne de donnée temp1∗ray1 obtenue par multiplication terme à terme, et on relance GAM à partir des nouvelles données. Ceci revient à chercher la fonction f sous la forme

$$P = f(\text{temp1}, \text{temp2}, ..., \text{raystat2}, \text{temp1} * \text{temp2}).$$

On compare alors la qualité des prédictions que l'on obtient avec ces nouvelles données par rapport à celles obtenues avec les données initiales, et on en déduit alors si notre regroupement de paramètres était pertinent ou non.

La question est alors : comment regrouper ces paramètres ? En effet, même si on ne garde que les cinq variables temp5, ray3, nebu0, raystat1 et raystat2 que l'on a déclaré comme étant les variables prépondérantes du modèle, les possibilités de regroupement sont déjà infinies : on a $2^5 = 32$ associations possibles, et encore ensuite une infinité d'opérateurs mathématiques : doit-on considérer un produit, une somme, une exponentielle, ... ?

Une méthode qui pourrait être valable ici mais qui ne s'exporte pas du tout dans un contexte plus général est d'étudier un modèle physique pour détecter des associations de paramètres. C'est toutefois une piste que nous n'avons pas eu le temps d'explorer davantage.

5 Modèle MARS

5.1 Présentation du modèle

Pour énoncer clairement les motivations de cette partie nous nous sommes référé à Friedman [1].
Un modèle MARS s'écrit sous la forme

$$\widehat{f}(x) = \sum_{m=1}^{M} a_m \; B_m(x), \tag{4}$$

avec $(a_m)_{1 \leq m \leq M}$ sont les coefficients d'ajustement du modèle et $(B_m)_{1 \leq m \leq M}$ sont les fonctions de base données par

$$B_m(x) = \prod_{k=1}^{K_m} \left[s_{km}.(x_{\nu(k,m)} - t_{km}) \right]_+ \tag{5}$$

La quantité s_{km} prend les valeurs ± 1, $\nu(k, m)$ désignent les indices des variables prédictrices et t_{km} sont la valeurs prises par ces variables. La conséquence d'appliquer les algorithmes de MARS (voir [1]) donne un modèle de la forme suivante

$$\widehat{f}(x) = a_0 + \sum_{K_m=1} f_i(x_i) + \sum_{K_m=2} f_{ij}(x_i, x_j) + \sum_{k_m=3} f_{ijk}(x_i, x_j, x_k) + ... \tag{6}$$

Soit $V(m) = \{\nu(k, m)\}_1^{K_m}$ l'ensemble des variables associées à la $m^{ème}$ fonction de base B_m. Alors

$$f_i(x_i) = \sum_{K_m=1,\ i \in V(m)} a_m \; B_m(x_i),$$

qui est la somme de toutes les fonctions de base impliquant seulement la variable x_i.

$$f_{ij}(x_i, x_j) = \sum_{K_m=2,\ (i,j)\in V(m)} a_m\ B_m(x_i, x_j),$$

qui présente la somme des fonctions de base à deux variables comportant notament la paire (x_i, x_j), de même on définit f_{ijk}.
La décomposition présentée dans (6) s'appelle le modèle ANOVA. Nous remarquons bien que cette décomposition donne la meilleure approche du modèle en question exprimé dans (3).
La projection de cet aspect théorique dans notre modèle physique a donné les résultats suivants.

5.2 Résultats obtenus par matlab et interprétations

– Création du modèle MARS.

Pour tester la validité du modéle créé on a utilisé la méthode de validation croisée, donc pour ce faire on divise k fois l'échantillon, puis on sélectionne un des k échantillons comme ensemble de validation et les $(k-1)$ autres échantillons constitueront l'ensemble d'apprentissage. On calcule comme dans la première méthode l'erreur quadratique moyenne. Puis on répète l'opération en sélectionnant un autre échantillon de validation parmi les $(k-1)$ échantillons qui n'ont pas encore été utilisés pour la validation du modèle. L'opération se répète ainsi k fois pour qu'en fin de compte chaque sous-échantillon ait été utilisé exactement une fois comme ensemble de validation. La moyenne des k erreurs quadratiques moyennes est enfin calculée pour estimer l'erreur de prédiction.

Variables	**Values**
$avgMSE$	0.0093
$avgRMSE$	0.0959
$avgRRMSE$	0.4008
$avgR2$	0.8384

avgMSE : Erreur moyenne quadratique
avgRMSE : Erreur moyenne quadratique relative
avgR2 : Coefficient de corrélation

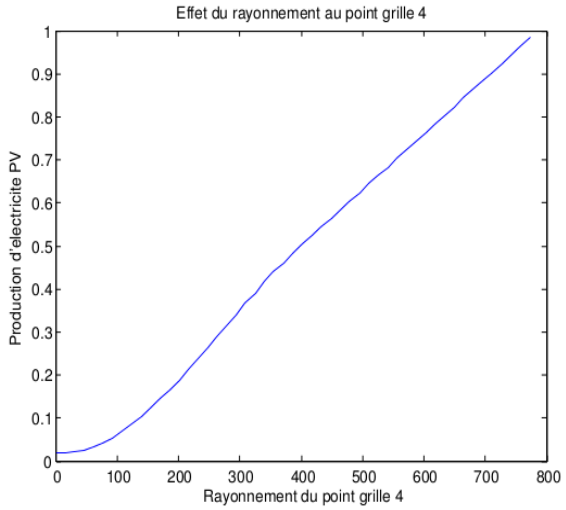
– Obtention des variables prédictrices associées aux différentes fonctions ANOVA :

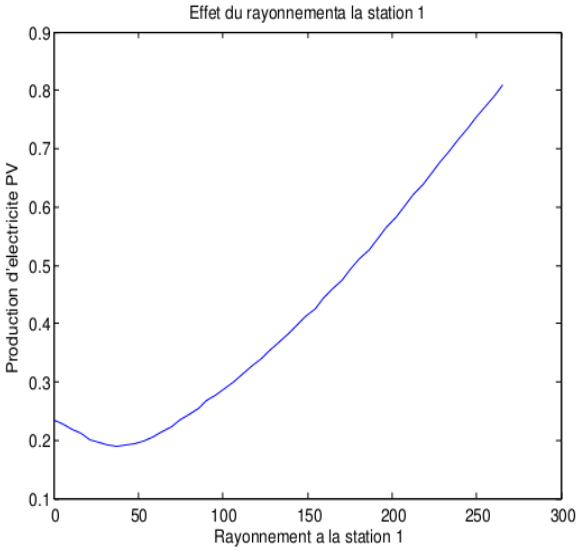
Fonction	**STD**	**GCV**	**Base**	**Variable**
1	0.191	0.058	1	$ray4$
2	0.112	0.089	2	$raystat1$
3	0.015	0.008	1	$temp2\ \ raystat1$
4	1.672	3.553	2	$temp2\ \ ray4$
5	1.639	3.421	2	$temp4\ \ ray4$
6	0.049	0.011	2	$vent4\ \ ray4$
7	0.091	0.046	2	$ray1\ \ raystat1$
8	0.063	0.013	2	$ray4\ \ nebustat0$
9	0.051	0.011	2	$nebustat0\ \ raystat1$

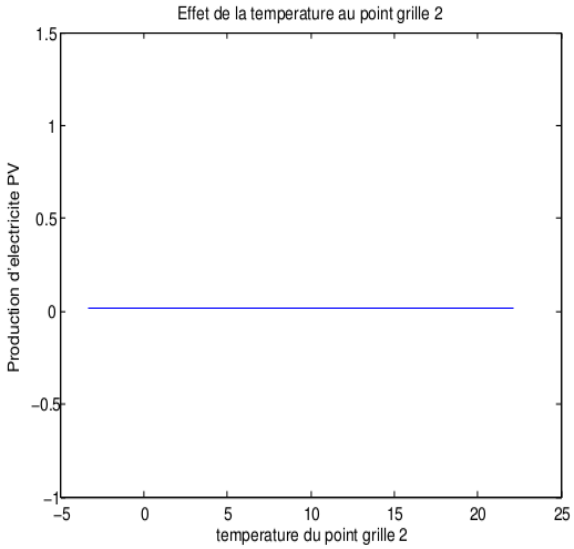
La décomposition ANOVA nous informe sur le nombre de fonctions qui on été utilisées pour la création du modéle, sur les variables prédictrices et aussi sur les variables predictrices qui sont en interaction. Donc grâce à la décomposition on sait que notre modéle s'écrit de la façon suivante :

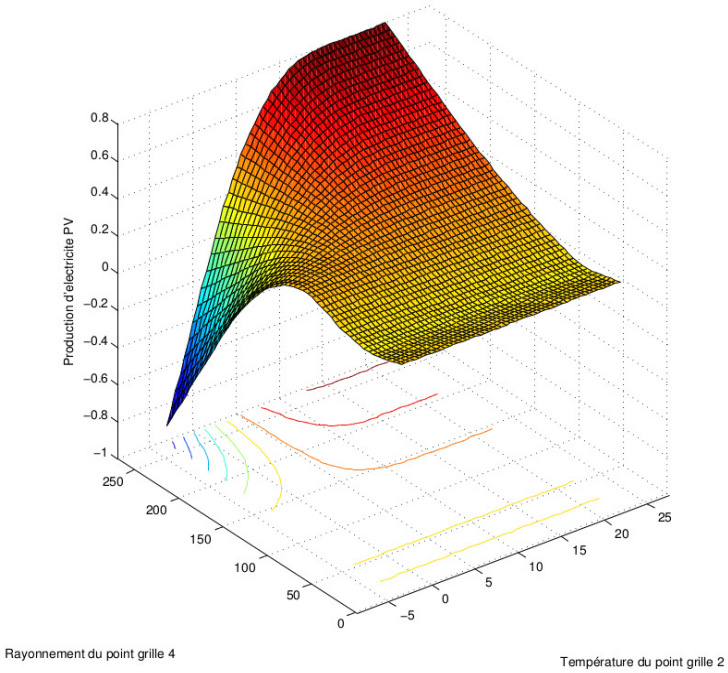
$$f = f_1(ray4) + f_2(raystat1) + f_3(temp2, raystat1) + ... + f_9(nebustat0, raystat1)$$

– Effet des variables prédictrices associées aux fonctions ANOVA :









6 Conclusion et perspectives

Grâce aux différentes méthodes que nous avons présentées, nous arrivons à prédire la production en J-1 et donner une corrélation entre les variables prédictrices. Cependant, la façon de regrouper les variables corrélées reste un problème, dont la résolution est nécessaire pour pouvoir adapter le modèle additif en un modèle plus proche de la situation étudiée.

Remerciements

Nous remercions RTE et particulièrement Mme. Laurence Maillard-Teyssier et Mr. Vincent Lefieux pour nous avoir proposé ce sujet, ainsi que pour leurs remarques qui ont conduit à des améliorations significatives du rapport. Nous remercions aussi Mme. Aurélie Gueudin pour ses discussions utiles sur quelques interprétations statistiques.

Références

[1] Friedman, J. H. (1991), Multivariate adaptive regression splines. *Ann. Statist.* Vol 19, No 1. pp. 1-67.